Google Cloud

Google Cloud’s Approach to Trust in Artificial Intelligence

Marina Kaganovich, Rohan Kanungo, Heidi Hanssen
Office of the CISO, Google Cloud

Disclaimer: The content contained herein is correct as of November 2023 and represents the status quo as of the time it was written. Google Cloud's security policies and systems may

change going forward as we continually make improvements for our customers.

Introduction

AI responsibility has come to be associated with not only mitigating risks but also helping improve people’s lives and addressing social and scientific challenges. Some elements – such as the need to incorporate essential societal values, including privacy, fairness, and transparency – are widely agreed upon, although striking the right balance between them can be complex. We recognize that advanced technologies can raise important challenges that must be addressed clearly, thoughtfully, and affirmatively.

In this paper, we provide insight into Google Cloud’s approach to building enterprise-grade AI responsibly. We share an overview of how we approach AI data governance, privacy, security, and compliance when developing generative AI, which customers can interact with through the ertex AI platformV.

For context, as used throughout this paper, generative AI refers to using AI to create new content such as text, images, music, audio, and video. Generative AI uses a machine learning (ML) model to learn the patterns and relationships in a dataset of human-created content. It then uses the learned patterns to generate new content. Generative AI is powered by foundation models (large AI models) that can multi-task and perform out-of-the-box tasks, including summarization, Q&A, classification, and more.

The Vertex AI platform is a machine learning platform that enables you to train and deploy machine learning models and AI applications and customize large language models (LLMs, a form of foundation models) for use in your AI-powered applications. Vertex AI combines data engineering, data science, and ML engineering workflows, enabling your teams to collaborate using a common toolset and scale your applications using the benefits of Google Cloud. In addition, you can also adapt foundation models for targeted use cases with minimal training and very little example data.

Google Cloud takes a principled approach to AI development, which is grounded in AI Principles that describe our commitment to developing technology responsibly and in a manner that is built for safety, enables accountability, and upholds high standards of scientific excellence. In furtherance of upholding these Principles, Google Cloud has identified specific areas we will not pursue, which include the design or deployment of technologies that are likely to cause overall harm or whose principal purpose or implementation is to cause or

directly facilitate injury to people. Likewise, we won’t pursue technologies whose purpose contravenes widely accepted principles of international law and human rights or whose use enables information gathering for surveillance that violates internationally accepted norms.

We’ve also infused these values into the oogle Cloud Platform Acceptable Use PolicyG and the Generative AI Prohibitive Use Policy so that they are transparent and communicated. In addition, when it comes to AI, we recognize the need for both good individual practices and shared industry standards. We’ve continued evolving our practices, conducting industry-leading research on AI impacts and risk management, and assessing proposals for new AI research and applications to ensure they align with our principles. We continuously iterate and reassess how to build accountability and safety into our work and publish our progress to encourage collaboration and advancements in this field.

As AI technologies advance, they provide the opportunity to enhance how we identify, address, and reduce security risks. We’ve taken a three-pronged approach to secure, scale, and evolve the security ecosystem by:

- Supporting customers in their AI implementation with controls, best practices, and capabilities
- Continuing to launch cutting-edge, AI-powered products and services to help organizations achieve better security outcomes at scale
- Continuously evolving to stay ahead of threats

Responsible Innovation

As with any transformational and new technology, we understand that AI comes with complexities and risks, and these need to be managed as part of a comprehensive risk management framework and governance structure. We recognize that AI presents critical questions, and we are working to build AI responsibly to benefit both our customers and the broader societies in which we operate. The challenge is to do so in a way that is proportionately tailored to mitigate risks and promote reliable, robust, and trustworthy AI applications while still enabling innovation and the promise of AI for societal benefit.

As part of our principled approach to building AI technologies, we commit to developing and applying strong safety and security practices and incorporating our privacy principles in the development and use of AI. Rigorous evaluations are a critical component of building successful AI. In Google Cloud, a large number of teams are engaged in the analyses and risk assessments for the AI products we build and early-stage customer co-development opportunities involving a customized approach.

Responsible product innovation spans multiple dimensions – some are technical, involving evaluations of data sets and models for bias; some pertain to product experiences, and some are around policy and are informed by decisions about what we will and won’t offer from a policy perspective. We’ve developed a process to review projects against the AI Principles and work with subject matter experts on privacy, security, and compliance, to name a few. We also regularly publish the progress we’re making to enable transparency in our work, support safer and more accountable products, earn and keep our customers’trust, and foster a culture of responsible innovation.

AI Risk Assessments

We’ve built an industry-leading governance process to align AI projects across the company with our AI Principles, providing education and resources, as well as structures and processes, including risk assessment frameworks through which we evaluate and guide our development and use of AI products and services.

During reviews, potential risks posed by the AI product being developed are identified and assessed. Strategies to mitigate potential harmful impacts are developed using our AI Principles as a guiding framework, as well as external frameworks and emerging best practices. We take a socio-technical approach, considering how AI will interact with the world and existing social systems. We assess the potential impacts and risks that maybe posed both at the initial release and as time goes on. Where appropriate, we develop mitigation strategies to address potential risks that may be identified prior to releasing the product for general availability (GA).

Mitigations can take various forms. For instance, for Generative AI products, mitigations may draw on technical approaches to evaluate and improve models during development, establish policy-driven safety guardrails, or may be enabled by tooling customers can leverage in their projects for further safety efforts. They may also be a combination of any of the above as best suited for the particular product being assessed, and are often informed by customer feedback. Policy restrictions are typically guided by the relevant Acceptable Use Policy, Terms of Service, and privacy restrictions, as further discussed in the AI Data Governance and Privacy section below.

Technical controls include using guardrails such as safety filters, which can block model responses that violate policy guidelines, for instance, around child safety. Model responses are, by default, blocked based on the probability that they contain violative content that falls within a defined set of“safety attributes.”These safety attributes also include“harmful categories” and topics that maybe considered sensitive, such as“drugs”or“derogatory.”For most of these categories, customers can set their thresholds for blocked responses and control content based on their business needs. Leveraging citation metadata is another example of a technical guardrail supporting the responsible use of AI. It can help manage the risk of Generative AI replicating existing content at length, and in those instances where a quote beyond a defined character threshold is sourced from a particular webpage, includes a citation to that page.

These controls can be further enhanced by tooling customers can use to achieve greater understanding and control of their AI models. For instance, leveraging models similar to those used in our safety filters, customers can use our text moderation service to scan the entire corpus of training set data for terms that fall within predefined“harmful categories”and topics that may be considered sensitive, enabling ongoing compliance through the identification of content that may violate relevant policies. Our explainable AI offerings provide tools and frameworks to help customers understand and interpret predictions made by machine learning models natively integrated with several Google Cloud products and services. Additionally, model evaluation on Vertex AI includes metrics to understand model performance and evaluate potential bias using common data and model bias metrics. These tools can promote fairness by evaluating data and model outputs during training and overtime, highlighting areas of concern and providing suggestions for remediation.

AI Data Governance & Privacy

In addition to reviewing AI products and services for adherence to our AI Principles, as discussed above, our teams assess products for compliance with data privacy and transparency requirements. These reviews also consider adherence to the commitmentsmade to our customers regarding data privacy and protection – specifically, the ability to control how customer data is accessed, used and processed – as articulated in the oogle Cloud PlatformG

erms of ServiceT and loud Data Processing AddendumC. In addition, we provide customers with visibility into who can access their data and why.

Our approach includes incorporating privacy design principles, designing architectures with privacy safeguards, and providing appropriate transparency and control over the use of data. When bringing new offerings to the market, we incorporate these principles throughout the product lifecycle and design architectures with comprehensive privacy safeguards such as data encryption and the ability to turn relevant features on or off.

We have also implemented robust data governance reviews to ensure data privacy commitments are considered when developing and deploying its AI products. One of the questions we are frequently asked is whether our foundation models are trained on customer data and, by extension, whether customer data may, as a result, be exposed to Google Cloud, Google Cloud's other customers, or the public. To address this question, we outline some key aspects of our model tuning and deployment, data governance practices, and our approach to privacy in our Vertex AI oferings below. This paper ofers additional information regarding

foundation model adaptation.

- Google Cloud processes customer data to provide our services. Google Cloud does not use customer data to train our foundation models without the customer's permission or instruction.

- The foundation models on Vertex AI are developed to handle general use cases. Customers can customize foundation models for speciic use cases using our tuning APIs. This approach combines our research and product development expertise to enable world-class AI without privacy compromises.

- Vertex AI ofers a parameter emcient tuning service that enables tuning of the foundation model to speciic tasks without having to rebuild the entire model. Each tuning job results in the creation of a few additional learned parameters, called "adapter weights," that are outside the foundation model. The adapter weights are speciic to the customer and only available to those who tuned those weights. During inference, the foundation model receives the adapter weights, runs through the request and returns the results without modifying the foundation model or storing the request.

- Input data, including prompts and adapter weights, are considered customer data and are stored securely at every step along the way - encrypted at rest and in transit. Customers can control the encryption of the stored adapter weights by using customer-managed encryption keys (CMEK) and can delete their adapter weights at any time. Customer data used to train adapter models will not be logged or used for improving the foundation model without the customer'spermission.

We are commited to preserving our customers' privacy with our Cloud AI oferings and supporting their compliance journey. As a global cloud provider, Google Cloud is commited to GDPR compliance and has compiled comprehensive DPIA Resource Center documentation to support customers in their data protection impact assessment eforts. We also enable certain AI/ML services to be conigured to meet data residency

requirements, as noted in our ervice TermsS.

AI Security

Privacy is tightly intertwined with security, and both are primary design criteria for all products built on Google Cloud. AI products are no exception, benefitting from Google Cloud's globally distributed and redundant infrastructure and inheriting the platform's foundational controls. Layered security controls protect AI models running on Google Cloud as we don't rely on any single technology to secure our infrastructure. Instead, our technology stack builds security through progressive layers that deliver defense in depth.

Our AI products are built atop a scalable technical infrastructure designed for maximized availability and reliability while providing security through the entire information processing lifecycle. Google Cloud's core principles include defense in depth, at scale, and by default. Data and systems are protected through multiple layered defenses using policies and controls configured across Identity and access management (IAM), encryption, networking, detection, logging, and monitoring. The platform is further underpinned by a secure-by-design foundation supported by operational controls consisting of in-depth security reviews, vulnerability scanning, ongoing threat monitoring, and intrusion detection mechanisms that enable secure service deployment and safeguard customer data. The security controls specific to Vertex AI, and generative AI features can be found here.

We strive to enable the security of our AI products and support customers in using AI to bolster their security capabilities. For instance, ecurity AI WorkbenchS is an industry-first extensible platform powered by Sec-PaLM 2, a specialized security large language model. This model is fine-tuned for security use cases, incorporating our expert security expertise, such as our visibility into the threat landscape and Mandiant's frontline intelligence. Sec-Palm 2 powers new offerings that can now uniquely address three top security challenges: threat overload, toilsome tools, and the talent gap, providing defenders with more natural, creative, andeffective ways to keep their organizations safe.

Our AI advances can now combine world-class threat intelligence with point-in-time incident analysis and threat detections and analytics to help prevent new infections, make security more understandable while helping to improve its effectiveness, and reduce the number of tools organizations need to secure their vast attack surface areas and ultimately, empower systems to secure themselves. Customers using Vertex AI can also benefit from ensitive DataS

Protection, which enables the identification of sensitive data such as email addresses, phone numbers, job titles, etc. based on pattern or a list, and then automatically hides or transforms that data by using methods such as masking or tokenization. This tool can also redact sensitive data, such as social security numbers, from images before ingesting it into a machine learning training environment.

Secure AI Framework (SAIF)

In addition to building our AI products on a secure platform, we’re committed to securely developing them using the ecure AI Framework (SAIF)S, a conceptual framework for securing AI systems. It’s inspired by the security best practices – like reviewing, testing and controlling the supply chain – that we apply to software development while incorporating our understanding of security megatrends and risks specific to AI systems. SAIF offers a practical approach to address the concerns that are top of mind for customers, including security, AI/ML model risk management, privacy, compliance and others. Customers may wish to considerSAIF as they define and refine their approach to adopting AI.

A framework like SAIF, which spans the public and private sectors, is essential for safeguarding the technology that supports AI advancements so that when AI models are implemented, they’re secure by default. The six core elements of SAIF are summarized below:

1. Expand strong security foundations to the AI ecosystem
2. Extend detection and response to bring AI into an organization’sthreat model
3. Automate defenses to keep pace with existing and new threats
4. Harmonize platform-level controls to ensure consistent security across the organization
5. Adapt controls to adjust mitigations and create faster feedback loops for AI deployment
6. Contextualize AI system risks in surrounding business processes

These steps aren’t simply conceptual. We’re putting them into action to support and advance a framework that works for all. It’s also important to consider that while there are novel aspects to securing AI, many current approaches to developing, deploying, and utilizing AI systems can be adjusted to account for these specificities rather than requiring a completely new approach. In this paper, we explore in greater depth what changes and what stays the same regarding AI cybersecurity.

AI Compliance

Security and privacy in cloud computing are topics often subject to legal, regulatory compliance, and risk management requirements. This is often the case in regulated industries such as financial services and healthcare and for certain critical service providers. Organizations running workloads and storing data on Google Cloud rightfully seek assurances as to the platform's control posture, frequently requiring documentation from an independent third party to validate their existence and efficacy. To support these expectations and provide transparency into its controls, Google Cloud makes compliance documentation, certifications, control attestations, and independent audit reports readily available to satisfy regional and industry-specific requirements and support customers in their compliance validation efforts of Google Cloud's platform, as well as their assessment of Vertex AI's compliance and security controls.

Similar to the dynamic nature of regulatory guidance promulgated as cloud computing transitioned from a developing to mainstream technology, we see a similar pattern emerging concerning AI regulation. The rapid advance of AI has captured regulators'attention worldwide, who are increasingly interested in understanding how current regulatory frameworks address AI and what new measures might be necessary to ensure AI is developed and deployed in a way that respects laws, norms, and human rights. This pace is exemplified in Stanford University's 2023 AI Index, which shows 37 AI-related bills were passed into law globally in 2022 alone.

We believe that AI is too important not to regulate and too important not to regulate well, and thus advocate for risk-based frameworks that reflect the complexity of the AI ecosystem by building on existing general concepts. We previously published recommendations for regulating AI, which outlined a general approach and some key implementation practicalities for policymakers to consider in developing practical AI regulations. Our top-line recommendations included:

- Taking a sectoral approach that builds on existing regulation
- Adopting a proportionate, risk-based framework
- Promoting an interoperable approach to AI standards and governance
- Ensuring parity in expectations between non-AI and AI systems
- Recognizing that transparency is a means to an end

Since then, we've further clarified our position, publishing A Policy Agenda for Responsible rogress in Artificial IntelligenceP and pplying model risk management guidance to artificialA intelligence/ machine learning-based risk models. In addition, our teams have adopted a risk assessment process to help:

1. Identify, measure, and analyze ethical risks throughout the life of an AI-powered product
2. Map these risks to appropriate mitigations
3. Develop clearer standards of acceptable risk

This process now also draws upon the best practices of various teams and is aligned with upcoming regulatory requirements in the U.S. and E.U.

We also closely track and monitor forthcoming industry standards such as the ISO/IEC 42001 AI Management System Standard, the recently published NIST AI Risk Management Framework, and global regulatory developments to ensure we continue to develop and deliver tools that serve our customers' needs. Google Cloud is a trusted voice in the international and regional standards development community. We actively provide feedback and shape the regulations, standards, and framework.

AI Environmental Impact

AI models and services can consume vast amounts of energy, which raises the responsibility for managing the carbon footprint resulting from the computing power required to train and run foundation models. As part of our ongoing work, we have identified four best practices that reduce energy and carbon emissions significantly, which we refer to as the "4Ms," which are being used today and are available to anyone using Google Cloud services. These four practices, each of which is briefly noted below, can, when implemented together, reduce energy by 100x and emissions by 1000x.

- **Model*.*** Selecting efficient ML model architectures, such as sparse models, can advance ML quality while reducing computation by 3x-to-10x.

- **Machine*.*** Using processors and systems optimized for ML training versus general-purpose processors can improve performance and energy efficiency by 2x-to-5x.

- **Mechanization.** Computing in the cloud rather than on-premises typically reduces energy usage as cloud-based data centers are newer, custom-designed warehouses generally equipped for energy efficiency. On-premises datacenters are often older and smaller and thus cannot amortize the cost of new energy-efficient cooling and power distribution systems.

- **Map Optimization.** The cloud enables customers to pick the location with the cleanest energy, reducing the gross carbon footprint by 5x-to-10x.

To minimize the potential for such environmental impact, Google Cloud has invested in developing efficient data centers. The cloud supports many products simultaneously, so it can more efficiently distribute resources among many users. That means we can do more with less energy. We're constantly looking for new ways to build products, design out waste and pollution, and keep materials and resources in use for as long as possible. We aim to maximize the reuse of finite resources across our operations, products, and supply chains and to enable others to do the same.

We're committed to operating carbon-free by 2030 and replenishing 120% of the water we consume by 2030. We're also dedicated to raising our standard of water stewardship, improving water quality and security, and restoring the health of ecosystems in the communities in which we operate.

Conclusion

We aim to be at the forefront of advancing AI through our deep research to develop a more capable and helpful AI. We're pursuing innovations to help unlock scientific discoveries and tackle some of humanity's most significant challenges. From this research and development, we are bringing innovations into the world to assist people and benefit society everywhere through our infrastructure, tools, products, and services. We are also enabling and working with others to benefit society.

To further the dialogue, we publish educational content, research and other forms of documentation to enable transparency and support our customers. These include Responsible AI Guides with best practices to assist customers in definingAI use cases and assessing their impact and AI research on various topics, including machine intelligence, natural language processing, and many others that maximize both scientific and real-world impact.

earning pathsL provide an overview of generative AI concepts, from the fundamentals of large language models to responsible AI principles. In addition, our recommended practices forAI are a helpful guide to follow when designing, developing and testing AI systems with a focus on fairness, interpretability, privacy, safety and security, including relevant examples and documentation for implementing each.

For further reading, consider the following resources:

- ecuring AI: Similar or DifferentS
- eople + AI GuidebookP for a set of methods, best practices and examples for designing with AI
- Built with Google Cloud AI for solutions and partners to support and accelerate AI implementation
- verview of Generative AI support on VertexO AI
- enerative AI on Google CloudG
- enerative AI Learning PathsG
- ntroduction to Vertex ExplainableI AI
- oogle Cloud Compliance Resource CenterG
- oogle Cloud Privacy Resource CenterG
- olicy Agenda for Responsible Progress in Artificial IntelligenceAP
- pplying model risk management guidance to artificial intelligence/ machine learningA based risk models